SOFE 4630 Winter 202 - Cloud Computing

Project Milestone#3 : Data Storage Implementation: KV + relational

Course Group No: 12

Group Members:

Ashwin Shanmugam - 100700236
Faazil Shaikh - 100707829
Evans Mosomi - 100719552
Samuel Rantetoding - 100694161

1. **Watch the first three videos for Kafka connectors (focus on the concepts, not the details) from https://www.confluent.io/blog/kafka-connect-tutorial/.**

2. **Describe the following:**
    - **Sink and Source connectors.**
    - **The applications/advantages of using Kafka Connectors with data storage.**
    - **How do Kafka connectors maintain availability?**
    - **List the popular Kafka converters for values and the properties/advantages of each.**

- A sink connector conveys information from Kafka topics into different frameworks, which may be records.
- A source connector retrieves data from a system. It can be the whole database, a stream table, or even a message broker
- A source connector could also gather measurements from application servers into Kafka topics, making the information accessible for stream handling with low latency.
- The advantages include reusability as we can use existing connector implementations for common data sources and sinks or implement our own connectors.
- Kafka connectors maintain availability by providing the ability to deploy the system as a standalone application or as a distributed application.
- JSON schema converter, it is used to read or consume JSON data from kafka topics into a Kafka Connect sink.
- Avro converter, It is used to convert data for Kafka Connect to and from Avro format.
- Protobuf is similar to XML but faster, the converter works with the registry to convert the data in and out of the format

3. **Search the internet to answer the following question:**
    - **What's a Key-Value (KV) database?**
    - **What are KV databases' advantages and disadvantages?**
    - **List some popular KV databases.**

- Key-value database is a sort of nonrelational data set that utilizes a basic key-value technique to store information.
- It stores information as an assortment of key-value sets in which a vital fills in as an unique identifier. Both keys and values can be anything, going from straightforward objects to complex compound objects.
- The advantages:
    - Simplicity
    - Scalability
    - Faster application response
- The disadvantages:
    - Optimized only for data with a single key and value. A parser is required to store multiple values.

    - Not optimized for lookup as it requires scanning the whole collection or creating separate index values.
- Popular KV databases:
    - Amazon DynamoDB
    - Aerospike
    - Redis
    - Berkeley DB

4. **Follow the following videos to deploy and use Redis and MySQL databases using GKE.**
    - Video Demo link: https://drive.google.com/file/d/1VCeUzmrZVNVfION6t-bY9KE8kPziKBK7/view?usp=sharing

5. **Follow the following video to set up sink and source Kafka connectors to the deployed MySQL database.**
    - Video Demo link: https://drive.google.com/file/d/1X2m3khf7CtZEG8PNw--LHpHVaf5QCLhi/view?usp=sharing

6. **Follow the following video to set up a Kafka connector to the deployed Redis database.**
    - Video Demo link: https://drive.google.com/file/d/1mHphKbXyHsxjDJrYE0ZUiwmJSBDWRZ3a/view?usp=sharing

7. **Now, you will store a dataset into cloud storage. The dataset has to be sent into Kafka topics and connectors have to be configured to automatically store the dataset into the data storage. The producer that will send the dataset to Kafka topics should run on your local machine as it will simulate real sensors while Kafka, connectors, and data storage should be on the cloud. Use MySQL for the CSV files and Redis for images. Feel free to update the Yaml files from the given repository to fit your dataset.**

8. **Use the sensors, images, ground truth Pose in the lastest session in http://robots.engin.umich.edu/nclt/ (Links to an external site.) as your dataset. Record a video showing the configuration of Kafka connectors, producers' python script, a proof of successfully stored data into data storage.**

9. **List some possible applications that can be implemented by using the uploaded dataset.**
    - Hydrodynamic Modeling
    - Geo-mapping
    - Object Detection

- Coastal Vulnerability Analysis
- Airborne Laser Swath Mapping

10. Upload the used files, reports, and videos (links) into your repository.

- Repository Link: https://github.com/ashwin1609/CloudComputing-Milestone3